ADAPTIVE CHOICE OF BOOTSTRAP SAMPLE SIZES

FRIEDRICH GÖTZE[1] AND ALFREDAS RAČKAUSKAS[2]

University of Bielefeld and University of Vilnius

Consider sequences of statistics $T_n(\widehat{P}_n, P)$ of a sample of size n and the underlying distribution. We analyze a simple data–based procedure proposed by Bickel, Götze and van Zwet (personal communication) to select the sample size $m = m_n \le n$ for the bootstrap sample of type "m out of n" such that the bootstrap sequence T^*_m for these statistics is consistent and the error is comparable to the minimal error in that selection knowing the distribution P. The procedure is based on minimizing the distance between $L_m(\widehat{P}_n)$ and $L_{m/2}(\widehat{P}_n)$, where $L_m(\widehat{P}_n)$ denotes the distribution of T^*_m.

Keywords and phrases: Bootstrap, m out of n bootstrap, Edgeworth expansions, model selection.

AMS subject classifications: 62D05.

1. Introduction

In this paper, we investigate an adaptive choice of the bootstrap sample size m in sampling from an i.i.d. sample of size n m-times independently and with (resp., without) replacement. To simplify the writing we shall abbreviate the notion of $\underline{m}$ $\underline{o}ut$ $\underline{o}f$ $\underline{n}$ sampling as *moon bootstrap*.

Assume that the random elements $X_1, \ldots, X_n, \ldots$ are independent and identically distributed from a distribution P on a measurable space $(S, \mathcal{A})$. Let $\widehat{P}_n$ denote the empirical measure of the first n observations $X_1, \ldots, X_n$. Throughout we assume that $P \in \mathcal{P}_o \subset \mathcal{P}$, where $\mathcal{P}_o$ is a set of probability measures on $(S, \mathcal{A})$ containing all empirical measures $\widehat{P}_n$.

Let $T_n = T_n(X_1, \ldots, X_n; P)$ denote a sequence of statistics, possibly dependent on the unknown distribution P in order to ensure that T_n is weakly convergent to some limiting distribution as n tends to infinity. A typical example is given by $T_n = n^\alpha\big(F(\widehat{P}_n) - F(P)\big)$, where $F : \mathcal{P}_0 \to R$ denotes a functional on $\mathcal{P}_0$ and $\alpha > 0$ is an appropriate normalization rate.

We are interested in the estimation of the distribution function (d.f.) $L_n(P; a)$ of T_n by means of resampling methods.

The nonparametric bootstrap estimates the d.f. $L_n(P, a)$ by the plug-in method, that is, by the conditional d.f.

$$\widehat{L}_n(a) := L_n(\widehat{P}_n, a) = P(T_n(X^*_1, \ldots, X^*_n; \widehat{P}_n) \le a | X_1, \ldots, X_n),$$

1 This research was supported by the Deutsche Forshungsgemeinshaft SFB 343.

2 This research was supported by the Deutsche Forshungsgemeinshaft SFB 343 and Institute of Mathematics and Informatics, Lithuania.

where $X_1^*, \ldots, X_n^*$ is a bootstrap sample from the empirical distribution $\widehat{P}_n$.

One of the major problems for the nonparametric bootstrap estimate $\widehat{L}_n$ is its consistency. Various types of consistency can be considered. Usually, if d denotes a certain distance on the set of all distribution functions then $\widehat{L}_n$ is said to be d-consistent (or, simply consistent, when d is fixed) in probability (resp., a.s.) provided that $d(L_n(P), \widehat{L}_n) \underset{n\to\infty}{\longrightarrow} 0$ in probability (resp., a.s.). Conditions ensuring the consistency were considered, e.g., by Bickel and Freedman (1981), Bretagnolle (1983), Athreya (1987) and Beran (1982, 1997). Extensive references and details on various bootstrap methods can be found in the recent monograph by Davison and Hinkley (1998).

A number of examples, where the bootstrap fails to be consistent together with positive results confirm that the consistency of the bootstrap estimate $\widehat{L}_n$ requires the following conditions:

1) for any Q from a neighborhood $\mathcal{V}(P)$ of P, $L_n(Q)$ has to converge weakly to a limit $L(Q)$, say, and the convergence has to be uniform on $\mathcal{V}(P)$;
2) the function $Q \to L(Q)$ has to be continuous.

The moon bootstrap with replacements (shortly, m/n-bootstrap) estimates the d.f. $L_n(P, a)$ by $L_m(\widehat{P}_n, a)$, whereas the moon bootstrap without replacements (shortly, $\binom{n}{m}$-bootstrap) estimates $L_n(P, a)$ by

$$L_m^*(\widehat{P}_n, a) = \binom{n}{m}^{-1} \sum_{1\le i_1<\cdots<i_m\le n} \mathrm{I}\{T_m(X_{i_1}, \ldots, X_{i_m}; \widehat{P}_n) \le a\}.$$

Under very weak conditions, the moon bootstrap resolves problems of consistency of the classical bootstrap by choosing $m = o(n)$ bootstrap samples. It was first suggested by Bretagnolle (1983) and investigated in Götze (1993), resp. Bickel, Götze and Zwet (1997) (examples of nonregular statistics), Politis and Romano (1994), Shao (1994), Swanepoel (1986), Deheuvels, Mason and Shorack (1990) (extreme value statistics), Shao (1996) (model selection), Datta and McCormik (1995), Datta (1996), Heimann and Kreiss (1996) (first order autoregression models), Athreya (1987), Arconess and Gine (1989) (heavy tailed population distributions). If d denotes a distance between distribution functions (e.g., Kolmogorv, Levy-Prokhorov, or bounded Lipschitz distance) then a measure of risk in estimating $L_n(P)$ by some estimator $\widehat{L}_n$ is given by

$$E_P d(\widehat{L}_n, L_n(P)).$$

For $\widehat{L}_n$ being the moon bootstrap estimator $L_m(\widehat{P}_n)$ the 'generic' nonparametric case is described by the nonconsistency of this estimator for $m \sim n$ due to the essential randomness of its limit distribution under $\widehat{P}_n$ for such m. Introducing $h := n/m$ as a 'bandwidth' type parameter in this nonparametric estimation problem, the case $h \sim 1$ is characterized by the fact that the variance of the bootstrap estimate may not tend to zero as n tends to infinity. On the other hand for large values of h the variance decreases in many cases of order $\mathcal{O}(h^{-1})$. Since the moon bootstrap actually estimates $L_m(P)$, the difference $d(L_m(P), L_n(P))$ will be significant for m/n small (or h large) and contributes a bias term which dominates the estimation error in this case. Thus, as in most nonparametric problems one has to look for a tradeoff choice of m minimizing the estimation error. On the other hand for 'parametric' problems where the bootstrap works like in the estimation of the distribution of Student's test statistic under the hypothesis one can show by higher order approximations that the bias as well as the variance of $L_m(\widehat{P}_n)$ essentially decrease as m grows up $m \sim n$, see Hall (1992). One would like to find a common recipe for choosing m effectively for both nonparametric as well as parametric situations in order to obtain a uniformly consistent and effective estimate for the distribution $L_n(P)$.

One way would be to look for a sample size m minimizing some cross-validation measure by a jackknife estimate related to the risk under the unknown distribution. This has been suggested by Datta and McCormik (1995) in a first order regression problem. Unfortunately, this method is computationally rather involved and the performance of this scheme is difficult to analyze.

Bickel, Götze and van Zwet (personal communication) suggested to base the choice of m on the discrepancy between $L_m(\widehat{P}_n)$ and $L_{m/2}(\widehat{P}_n)$. We motivate this idea by showing that the (random) distance between $L_m(\widehat{P}_n)$ and $L_n(P)$ as a function on m is *stochastically equivalent* to the (random) distance between $L_m(\widehat{P}_n)$ and $L_{m/2}(\widehat{P}_n)$ as $n \to \infty$ and $m = m(n) \to \infty$. More precisely, consider for some distance d between distributions (like e.g. Kolmogorov's distance)

$$\Delta_m := d\left(L_m(\widehat{P}_n), L_n(P)\right) \quad \text{and} \quad \widehat{\Delta}_m := d(L_m(\widehat{P}_n), L_{m/2}(\widehat{P}_n)). \tag{1.1}$$

In Theorem 3.3 we prove that, under certain conditions, for some model dependent rate $0 < \alpha < 1$ (like $= 1/2, 1/3, 1/4$ etc.) and sequences $m(n)$ such that the limit $\lim_n m(n)/n^\alpha$, say $\gamma \in [0, \infty]$ exists, we have

$$\frac{\widehat{\Delta}_m}{\Delta_m} \underset{n\to\infty}{\overset{\mathcal{D}}{\longrightarrow}} \xi_\gamma \tag{1.2}$$

where ξ_0 and ξ_∞ are constants depending on P and, for $0 < \gamma < \infty$, ξ_γ denotes a random variable depending on P. Here, $\overset{\mathcal{D}}{\longrightarrow}$ denotes as usual convergence in distribution. Typically, we find that $E_P\widehat{\Delta}_m/E_P\Delta_m \longrightarrow c_\gamma(P)$, where $c_\gamma(P)$ is a constant depending on γ and P.

Based on these observations, we suggest $m^* = \mathrm{argmin}_{2\le m\le n}\, \widehat{\Delta}_m$ as a random choice of m for the m/n-bootstrap. We will show that this choice is as good as choosing the optimal m when knowing the unknown distribution P as long as $m/n \to 0$ holds. Simulations show that the method works in the region $m \sim n$ as well but behavior in this region is difficult to analyze for general models of distributions.

The reasons for such a choice are illustrated by the following simple example.

Example 1.1. Consider the statistic $T_n = T_n(X_1, \ldots, X_n; P) = n(\overline{X}_n)^2$, where $X_1, \ldots, X_n$ is an i.i.d. sample from a distribution P on the real line with zero mean and let $\overline{X}_n$ denote the sample mean. Let $L_n(P;r)$ denote the d.f. of T_n. The corresponding m/n-bootstrap approximation is the d.f. $L_m(\widehat{P}_n;r)$ of the statistic $T_n^* = m(\overline{X_m^*})$, where $X_1^*,\dots,X_m^*$ is a sample from the empirical distribution $\widehat{P}_n$, and $\overline{X_m^*}$ denotes the corresponding sample mean. Assume that $EX_1^4 < \infty$ and P satisfies Cramér's condition of smoothness. Consider the uniform errors introduced in (1.1), based on the Kolmogorov distance d. Let Y denote a standard normal random variable and assume that the sequence $m = m(n)$ is chosen such that $m \to \infty$ and $m/n \to 0$. In Section 4 we prove the following. If $m/n^{1/2} \to \infty$, then

$$(n/m)\big(\Delta_m, \widehat{\Delta}_m\big) \underset{n\to\infty}{\overset{\mathcal{D}}{\longrightarrow}} (c_1 Y^2, c_1 Y^2/2), \tag{1.3}$$

where c_1 denotes an absolute constant. If $m/n^{1/2} \to 0$, then

$$m\Delta_m \overset{P}{\longrightarrow} c_0(P), \quad \text{and} \quad m\widehat{\Delta}_m \overset{P}{\longrightarrow} c_0(P), \tag{1.4}$$

where $c_0(P)$ is a constant depending on P. If $m/n^{1/2} \to c = const. \neq 0$, then

$$n^{1/2}\big(\Delta_m, \widehat{\Delta}_m\big) \underset{n\to\infty}{\overset{\mathcal{D}}{\longrightarrow}} (f_0(Y), f_1(Y)), \tag{1.5}$$

where f_0, f_1 are certain measurable functions of Y (see Section 4 for details). Thus, if $\lim m/n^{1/2} = \gamma \in [0,\infty]$ exists, then (1.3)–(1.5) imply (1.2). Moreover, $\xi_0 = 1$ and $\xi_\infty = 1/2$. Under the same conditions we obtain as well $E_P\widehat{\Delta}_m/E_P\Delta_m \underset{n\to\infty}{\longrightarrow} c_\gamma(P)$.

Note that the value of m obtained in this way by minimizing $\widehat{\Delta}_m$ strongly depends on the particular sample. For instance if by chance in this example $\overline{X_n}$ approximates the true value $EX_1 = 0$ very accurately, that is $n^{1/2}\overline{X_n} = o(1)$ (which happens rarely), the approximation of $L_n(P)$ by the random bootstrap distribution $L_n(\widehat{P}_n)$ might be accidentally precise as well (compare (4.1) and (4.2)). In this case the bias as well as the variance of the estimate $L_m(\widehat{P}_n)$ will decrease with m which leads to a choice of a large sample size $m \sim n$. Such a case cannot be detected by an average criterion for the choice of m like, say $E_P\, d\left(L_m(\widehat{P}_n), L_n(P)\right)$, which would lead to a much less adaptive and accurate choice of m for such an exceptional sample.

Similar arguments apply to the moon bootstrap without replacements. We will show under certain conditions (see Theorems 2.1, 2.3 and Remark 2.1) that, for L_2-distances of distribution functions, say d, the random distance $\Delta_m^* = d(L_m^*(\widehat{P}_n), L_n(P))$ is stochastically equivalent to $\widehat{\Delta_m^*} = d(L_m^*(\widehat{P}_n), L_{m/2}^*(\widehat{P}_n))$. More precisely, for some model dependent rate $0 < \alpha < 1$ and sequences $m(n)$ such that $\gamma = \lim_n m(n)/n^\alpha \in [0,\infty]$ exists we have

$$\frac{\widehat{\Delta_m^*}}{\Delta_m^*} \underset{n\to\infty}{\overset{\mathcal{D}}{\longrightarrow}} \eta_\gamma,$$

where η_0 is a constant depending on P and, for $0 < \gamma \le \infty$, η_γ is a random variable depending on P. Typically we get $E_P\widehat{\Delta_m^*}/E_P\Delta_m^* \to c_\gamma$ (see Remark 2.1). This motivates $m^* = \operatorname{argmin}\widehat{\Delta_m^*}$ as a random choice of m for the $\binom{n}{m}$-bootstrap.

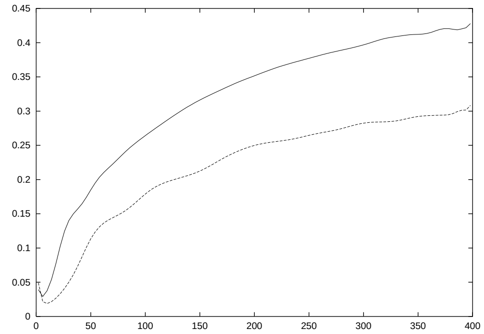

Fig.1.

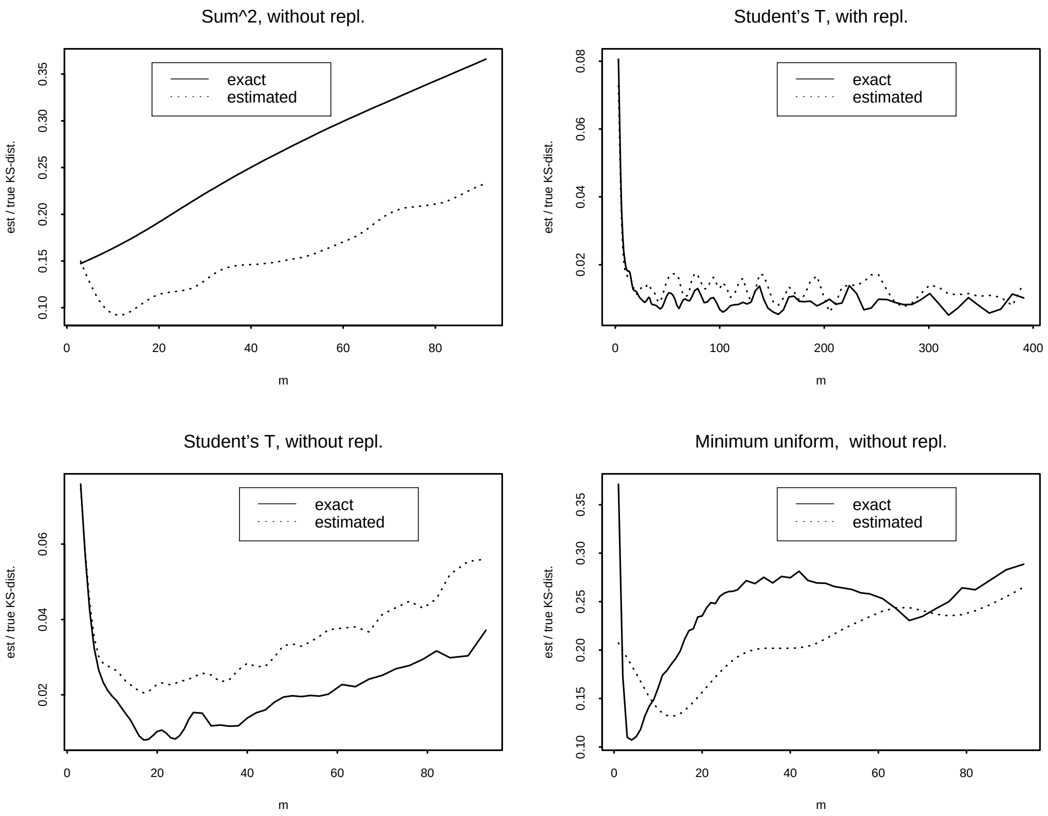

Fig.2.

In Fig.1 we consider Example 1.1. It shows (smoothed) graphs of the functions $m \to \widehat{\Delta}_m$ (dot line) and $m \to \Delta_m$ (solid line), where P is a centered χ_4^2 distribution. The simulations were done for a sample size of $n = 400$.

In Fig. 2 the true and estimated Kolmogorov distances $m \to \widehat{\Delta_m^*}$ and $m \to \widehat{\Delta_{2m}^*}$ are smoothed and plotted based on an individual sample of size $n = 400$ and $m \leq n/2$ when sampling without replacement. The first plot shows the behavior of sampling without replacement in the setup of Fig. 1 with $P = \chi_1^2$. The second and third plot represents a parametric case: Student's t-test with $P = N(0,1)$ and sampling with/without replacement. The third one represents a nonparametric case: distances for the normalized distribution of the largest order statistic for $P = \text{Uniform}(0,1)$ and sampling without replacement, see also Example 2.2 below.

The paper is organized as follows. In the Section 2 we investigate the moon bootstrap without replacements. To this aim Hoeffding expansions for U-statistics are used for $m/n = o(1)$ in order to evaluate the error of the random approximations. In Section 3 we investigate as well the moon bootstrap without replacements representing our statistics in terms of empirical processes. Here, following Beran (1994), we require that the sequence of statistics should be locally asymptotically weakly convergent. Furthermore, we shall use Edgeworth expansions to prove the stochastic equivalence of the random distances in the examples studied in this paper. Finally Section 4 contains the proofs of our results.

Throughout the paper we write $m \in n(\alpha, \gamma)$ to indicate that $m = m(n)$ is a sequence such that $m \to \infty$, $m/n \to 0$ and $\lim_n m/n^\alpha = \gamma$ exists allowing $\gamma \in [0, \infty]$.

Acknowledgements. The first author would like to thank Peter Bickel and Willem van Zwet for a number of stimulating and very helpful discussions on the construction of subsampling procedures.

2. Moon bootstrap without replacements

In this section, $X_1, \cdots, X_n$ denotes a sample of i.i.d. random elements from an unknown distribution P on a measurable space $(S, \mathcal{A})$, and $T_n = T_n(X_1, \ldots, X_n; P)$ will denote a sequence of statistics with distribution $L_n(P)$. We assume that T_n converges in distribution to a random variable T_∞.

Let $\theta_n(P; a) = Eh(T_n; a)$, where $h : R \times R \to R$ is a real measurable bounded function denote a family of parameters indexed by $a \in R$. The moon bootstrap without replacements estimates $\theta_n(P; a)$ by

$$\theta_{mn}(\widehat{P}_n; a) = L_m^*(\widehat{P}_n)h(\cdot, a) = \binom{n}{m}^{-1} \sum_{1 \le i_1 < \cdots < i_m \le n} h(T_m(X_{i_1}, \ldots, X_{i_m}; \widehat{P}_n); a).$$

As distance d between $L_n(P)$ and $L_m^*(\widehat{P}_n)$ we choose the L_2-distance between $\theta_n(P; a)$ and $\theta_{mn}(\widehat{P}_n; a)$ writing

$$\Delta_m^* = \Big(\int_R \big(\theta_n(P; a) - \theta_{mn}(\widehat{P}_n; a)\big)^2 \mu(da) \Big)^{1/2},$$

$$\widehat{\Delta_m^*} = \Big(\int_R \big(\theta_{mn}(\widehat{P}_n; a) - \theta_{Mn}(\widehat{P}_n; a)\big)^2 \mu(da) \Big)^{1/2}, \quad M = m/2,$$

where μ is a probability measure. For indicator functions $h(x; a) = \mathrm{I}\{x \le a\}$, Δ_m^{*2} reduces to the integrated square error between the distribution function $L_n(P; a)$ and its $\binom{n}{m}$-bootstrap estimator $L_m^*(\widehat{P}_n; a)$. For special discrete measures μ Δ_m^{*2} may be written as

$$\Delta_m^{*2} = \sum_k \lambda_k (L^*(\widehat{P}_n; a_k) - L_n(P; a_k))^2.$$

We shall give conditions which ensure the stochastic equivalence of Δ^*_m and $\widehat{\Delta^*_m}$. First, we impose some restrictions on the sequence of statistics T_n.

Assumption (I). There exist measurable functions $\kappa,\ \xi_m : S\times R\to R$ such that

$$E\left(h(T_m;a)\big|X_1\right)=Eh(T_\infty;a)+m^{-1/2}\kappa(X_1;a)+\xi_m(X_1;a),$$

where $\int_R E\ \kappa^2(X_1;a)\mu(da)<\infty$ and $\int_R E\ \xi_m^2(X_1;a)\mu(da)=o(m^{-1})$.

Assumption (J). $\int_R(\theta_{mn}(\widehat{P}_n;a)-\theta_{mn}(P;a))^2\mu(da)=o_P\big(m/n+1/m\big)$, where

$$\theta_{mn}(P;a)=\binom{n}{m}^{-1}\sum_{1\le i_1<\cdots<i_m\le n}h(T_m(X_{i_1},\ldots,X_{i_m};P);a).$$

To establish stochastic equivalence of Δ^*_m and $\widehat{\Delta^*_m}$ we shall consider stochastic processes $\zeta^*_{mn}(a)=\theta_{mn}(\widehat{P}_n;a)-\theta_n(P;a),\quad a\in R$, and $\widehat{\zeta^*_{mn}}(a)=\theta_{mn}(\widehat{P}_n;a)-\theta_{\frac{m}{2}n}(\widehat{P}_n;a),\quad a\in R$, as random elements in the space $L_2(R,\mu)$. Let ζ_P denote a mean zero Gaussian process with covariance given by

$$E\zeta_P(a)\zeta_P(b)=E\kappa(X_1;a)\kappa(X_1;b),\quad a,b\in R.$$

THEOREM 2.1. *Suppose that the assumptions (I) and (J) are satisfied and $\alpha_P(\cdot)=E\ \kappa(X_1;\cdot)\neq 0$. Choose sequence $m\in n(1/2,\gamma)$. Depeending on $\gamma=\infty$ or $0\le\gamma<\infty$ choose norming sequence $\tau_{nm}=(n/m)^{1/2}$ or $m^{1/2}$ respectively. Then*

$$\tau_{mn}(\zeta^*_{mn},\widehat{\zeta}^*_{mn})\ \xrightarrow[n\to\infty]{\mathcal{D}}\ (\xi_\gamma,\eta_\gamma) \tag{2.1}$$

in the space $L_2(R,\mu)\times L_2(R,\mu)$, where (writing $c:=1-2^{-1/2},\ d:=1-2^{1/2}$) $(\xi_\infty,\eta_\infty)=(1,c)\zeta_P$ and $(\xi_\gamma,\eta_\gamma)=(\gamma\zeta_P+\alpha_P,c\gamma\zeta_P+d\alpha_P)$ for $0\le\gamma<\infty$.

Noting that $\Delta^*_m=||\zeta^*_{mn}||_2$ and $\widehat{\Delta^*_m}=||\widehat{\zeta}^*_{nm}||_2$, where $||\cdot||_2$ denotes the norm in $L_2(\mu)$ we have by Theorem 2.1 the following corollary.

COROLLARY 2.2. *Suppose that the assumptions (I) and (J) are satisfied and the sequences $m=m(n)$ and τ_{nm} are choosed as in Theorem 2.1. Then*

$$\frac{\widehat{\Delta^*_m}}{\Delta^*_m}\ \xrightarrow[n\to\infty]{\mathcal{D}}\ \tau_\gamma, \tag{2.2}$$

where $\tau_\infty=1-2^{-1/2},\tau_0=2^{1/2}-1$ and τ_γ is a random variable for $0<\gamma<\infty$.

Remark 2.1. From the proof of Theorem 2.1 it is clear, that if assumption (J) is substituted by

$$\int_R E(\theta_{mn}(\widehat{P}_n;a)-\theta_{mn}(P;a))^2\mu(da)=o\big(m/n+1/m\big),$$

then, in addition to (2.2), we also have $E_P\widehat{\Delta_m^*}/E_P\Delta_m^* \underset{n\to\infty}{\longrightarrow} c_\gamma(P)$. A similar remark applies to other results in this section. Furthermore, similar results hold with different $c = c(\lambda)$ when comparing sample sizes m and λm, $0 < \lambda < 1$ in $\widehat{\Delta_m^*}$.

In the following we shall discuss the nature of the assumptions of Theorem 2.1. Assumption (J) allows to reduce the analysis of the $\binom{n}{m}$-bootstrap approximation $\theta_{mn}(\widehat{P}_n)$ to that of U-statistics $\theta_{mn}(P)$ with increasing degree $m = m(n)$ and values in the space $L_2(\mu)$. This assumption is satisfied for a large class of statistics T_n. For example, we consider the estimation of a parameter $\theta(P)$ of an unknown distribution P by means of a plug-in estimator $\theta(\widehat{P}_n)$. Introduce the statistic $T_n = \tau_n(\theta(\widehat{P}_n) - \theta(P))$, where the normalization τ_n is such chosen such that T_n converges in distribution. If h is Lipschitz function, e.g., $\int(h(t;a) - h(s;a))^2 d\mu| \le |t-s|^2$, we obtain

$$\int (\theta_{mn}(\widehat{P}_n;a) - \theta_{mn}(P;a))^2 d\mu \le \tau_m|\theta(\widehat{P}_n) - \theta(P)|$$

and condition (I) is satisfied provided that $\tau_m/\tau_n = o\big((m/n)^{1/2}\big)$.

In case where $h(x;a) = \mathrm{I}\{x \le a\}$ is an indicator function, we obtain on the set $\tau_m|\theta(\widehat{P}_n) - \theta(P)| \le \varepsilon$, that $|\theta_{mn}(\widehat{P}_n) - \theta_{mn}(P)| \le U_{mn}h_\varepsilon$, where $U_{mn}h_\varepsilon$ is a U-statistic with kernel $h_\varepsilon(X_1,\dots,X_m) = \mathrm{I}\{a-\varepsilon \le T_m(X_1,\dots,X_m;P) \le a+\varepsilon\}$. Hence, $E|\theta_{mn}(\widehat{P}_n) - \theta_{mn}(P)|$ does not exceed the quantity

$$\inf_{\varepsilon>0}\Big\{P\big(a-\varepsilon \le T_m \le a+\varepsilon\big) + P\big(\tau_m|\theta(\widehat{P}_n) - \theta(P)| \ge \varepsilon\big)\Big\}$$

which, under the convergence assumption, is typically, of the required order.

Edgeworth expansions are well suited to estimating the accuracy of a bootstrap approximation (see, e.g., the monograph by Hall (1992)). Assumption (I) requires certain expansions for the distribution $L_m(P)$ as well as the conditional distribution $L_m(P)\big|X_1$. Moreover, by conditioning the statistic T_m on a random sample, we implicitly assume some kind of stochastic expansion for the function T_m in the sense that condition (I) may be checked by studying the function $y \to g(y) = Eh(T_m(y,X_2,\dots,X_m;P);a)$. Consider, for example, an i.i.d. sample $X_1,\dots,X_n$ from a distribution P on R^d. Let $EX_1 = 0$. For a given symmetric $d\times d$ matrix Q, consider the statistic $T_n = nQ(\overline{X}_n)$, where Q denotes a quadratic form with $Q(x) = \langle x, Qx\rangle$ for $x \in R^d$, where $\langle\cdot,\cdot\rangle$ denotes the Euclidean scalar product. Set $\overline{X}_{n|1} = n^{-1}\sum_{k=2}^n X_k$ and $T_{n|1} = nQ(\overline{X}_{n|1})$. If the function h has two bounded derivatives with respect to the first argument, we have

$$E\left(h(T_m;a)\big|X_1\right) = Eh(T_{m|1};a) + 2Eh'(T_{m|1};a)\langle X_1, Q\overline{X}_{m|1}\rangle + R_m(X_1;a),$$

where $ER^2(X_1;a) \le cm^{-2}$ provided $E|X_1|^4 < \infty$. Since $m^{1/2}\overline{X}_{m|1}$ converges in distribution to a mean zero Gaussian random vector Y_P with $\operatorname{cov} Y_P = \operatorname{cov} X_1$, one needs expansions for $E\psi(m^{1/2}\overline{X}_{m|1})$ and $E\phi(m^{1/2}\overline{X}_{m|1})$, where $\psi(x) = h(\langle x, Qx\rangle;a)$,

$\phi(x) = 2h'(\langle x, Qx\rangle; a)\langle X_1, Qx\rangle$. If the function h is sufficiently smooth, the required expansions are well known (see, e.g., Götze (1985)).

If $h(x;a) = \mathrm{I}\{x \le a\}$, then, evidently,

$$P(T_m \le a \big| X_1) = P(T_{m|1} + \langle X_1, Q\overline{X}_{m|1}\rangle \le a - m^{-1}Q(X_1)),$$

and the problem reduces to prove expansions for the distribution of the quadratic polynomial $T_{m|1} + \langle x, Q\overline{X}_{m|1}\rangle$. Such expansion have been proved up to order $O(m^{-1})$ in Bentkus and Götze (1999).

From the proof of Theorem 2.1 it is clear that one needs to control the covariance of $Eh(T_m;a)\big|X_1$ and the correlation between $Eh(T_m;a)$ and $Eh(T_{m/2};b)$ for $a, b \in R$. It is also clear that if $\alpha_P = 0$, one needs to assume that a second order approximation for $Eh(T_m;a)$ holds. Consider, for example,

Assumption (I_1). There exist two functions $\psi_1, \psi_2 : R \times R \to R$, possibly dependent on P, such that for all $a, b \in R$

$$\begin{aligned} \operatorname{cov}\big(E\big(h(T_m;a)\big|X_1\big); E\big(h(T_m;b)\big|X_1\big)\big) &= \frac{1}{m}\big(\psi_1(a,b) + o(1)\big) \quad \text{and} \\ \operatorname{cov}\big(E\,\big(h(T_m;a)\big|X_1\big)\,; E\,\big(h(T_{m/2};b)\big|X_1\big)\,\big) &= \frac{1}{m}\big(\psi_2(a,b) + o(1)\big). \end{aligned}$$

Assumption (I_2). $E_P h(T_m;a) = Eh(T_\infty;a) + m^{-1}\beta_P(a) + \delta_{P,m}(a)$, where $\beta_P \in L_2(\mu)$ and $\delta_{P,m}$ is $o(m^{-1})$ in $L_2(\mu)$.

Assumption (J'). $\int(\theta_{mn}(\widehat{P}_n;a) - \theta_{mn}(P;a))^2\mu(da) = o_P\big(m/n + m^{-2}\big)$.

Let ζ_1, ζ_2 denote mean zero Gaussian processes with covariances equal to ψ_1 and correlation $E\zeta_1(a)\zeta_2(b) = 2^{-1/2}\psi_2(a,b),\ a, b \in R$.

THEOREM 2.3. *Suppose that the assumptions $(I_1), (I_2)$, and (J') are satisfied and $\beta_P \ne 0$. Choose sequence $m \in n(1/3, \gamma)$. Depending on $\gamma = \infty$ and $0 \le \gamma < \infty$ we may choose norming sequences $\tau_{mn} = (n/m)^{1/2}$ and m respectively such that* (2.1) *holds where (writing $c_0 = 2^{-1/2}$)*

$$\begin{aligned} (\xi_\infty, \eta_\infty) &= \big(\zeta_1,\ \zeta_1 - c_0\,\zeta_2\big), \\ (\xi_\gamma, \eta_\gamma) &= (\gamma\zeta_1 + \beta_P, \gamma(\zeta_1 - c_0\,\zeta_2) - \beta_P) \quad \text{for} \quad 0 \le \gamma < \infty. \end{aligned}$$

Moreover, (2.2) holds where $\tau_0 = 1$ and τ_γ is a random variable when $0 \le \gamma < 1$.

Example 2.2. In this example, we assume that $X_1, \dots, X_n$ is a random sample from the uniform distribution P on the interval $(0, \theta)$. The maximum likelihood estimates θ by the extreme order statistic $\widehat{\theta} = \max_{1 \le k \le n} X_k$. Set

$$T_n := T_n(X_1, \dots, X_n, P) = \frac{n(\theta - \widehat{\theta})}{\theta}.$$

Let $L_n(P;a)$ denote the distribution function of this statistic. Its $\binom{n}{m}$-bootstrap approximation $L_m^*(\widehat{P}_n;a)$ is defined by

$$L_m^*(\widehat{P}_n;a) = \binom{n}{m}^{-1} \sum_{1\le k_1<\cdots<k_m\le n} \mathrm{I}\{T_m(X_{k_1},\ldots,X_{k_m};\widehat{P}_n) \le a\},$$

where

$$T_m(X_{k_1},\ldots,X_{k_m},\widehat{P}_n) = \frac{m(\widehat{\theta} - \max_{1\le i\le m} X_{k_i})}{\widehat{\theta}}.$$

Set

$$\widehat{E}_m^2 = \int_R \left(L_m^*(\widehat{P}_n;a) - L_{\frac{m}{2}}^*(\widehat{P}_n;a)\right)^2 \mu(da), \quad E_m^2 = \int_R \left(L_m^*(\widehat{P}_n;a) - L_n(P;a)\right)^2 \mu(da).$$

As a consequence of Theorem 2.3 we get in this example

Corollary 2.4. *Assume that $m \in n(1/3,\gamma)$. We have with the notations and choices of Theorem 2.3*

$$\frac{\widehat{E}_m}{E_m} \xrightarrow[n\to\infty]{\mathcal{D}} \tau_\gamma.$$

Proof. We shall verify the assumptions of Theorem 2.3, for $h(x;a) = \mathrm{I}\{x \le a\}$. Obviously we have, for $0 < a < n$,

$$\left(1 - \frac{a}{n}\right)^n = e^{-a} - e^{-a}\frac{a^2}{2n} + e^{-a}\frac{\theta a^3(1+a)}{n^2(1-a/n)^2}, \tag{2.3}$$

where $|\theta| \le 1$. Since T_n converges in distribution to an exponential random variable T_∞, the assumption (I_2') with $\beta_P(a) = -e^{-a}a^2/2$ follows by (2.3). By straightforward calculations we see that

$$E|L_m^*(\widehat{P}_n,a) - L_m^*(P,a)| \le \frac{m}{n}.$$

This, clearly, yields assumption (J'). To check assumption (I_1) note that for

$$\begin{aligned} Z_m(a) &:= P(T_m \le a\big|X_1) - P(T_m \le a) = \\ &= (1-a/m)^{m-1}\big[(1-a/m) - \mathrm{I}\{X_1 < \theta(1-a/m)\}\big], \end{aligned}$$

using (2.3) we have $EZ_m(a)Z_m(b) = \exp\{-(a+b)\}\min\{a;b\}m^{-1} + o(m^{-1})$ and $EZ_m(a)Z_{m/2}(b) = \exp\{-(a+b)\}\min\{a;2b\}m^{-1} + o(m^{-1})$. Hence assumption (I_1) follows which completes the proof of the corollary. ∎

3. Moon bootstrap with replacements

In the cases, where the classical nonparametric bootstrap fails for the distribution $L_n(P)$, it often turns out that the limit distribution of $L_n(\widehat{P}_n)$ is random. In order to describe it, we introduce, following Beran (1997), a local weak convergence property of the statistic which we want to approximate in distribution. In order to introduce this notion, we need some preparation.

Fix a set $\mathcal{F} \subset L_2(S,P)$. Let $\ell_\infty(\mathcal{F})$ denote the Banach space of real bounded functions on $\mathcal{F}$ equipped with the supremum norm $||z||_{\mathcal{F}} = \sup_{f\in\mathcal{F}} |z(f)|$. Assume that the envelope function

$$F(s) = \sup\{|f(s)| : f \in \mathcal{F}\} < \infty \quad \text{for all} \quad s \in S,$$

and the quantities

$$\sup\{|Qf| : f \in \mathcal{F}\} < \infty \quad \text{for all} \quad Q \in \mathcal{P}_0,$$

where $Qf = \int f dQ$ are bounded. Then the mappings $\delta_s : \mathcal{F} \to R$ given by $\delta_s f = f(s)$ and $Q : \mathcal{F} \to R$ given by $Qf = \int f dQ$, where $Q \in \mathcal{P}_0$, are in $\ell_\infty(\mathcal{F})$. In order to avoid measurability problems connected with the non-separability of $\ell_\infty(\mathcal{F})$, we assume throughout, that $\mathcal{F}$ is either a countable or an image admissible Suslin set. For the relevant definitions concerning these matters, we refer to Dudley (1984).

Let the empirical process $\nu_{P,n} = (\nu_{P,n}(f), f \in \mathcal{F})$ be defined by $\nu_{P,n} = n^{1/2}(\widehat{P}_n - P)$, $\; n \in N$. Note that, for each $Q \in \mathcal{P}_0$, $\nu_{Q,n}$ is a random element with values in $\ell_\infty(\mathcal{F})$.

Definition 3.1. The statistic $T_n = T_n(X_1,\ldots,X_n;P) \in R$ is called locally $\mathcal{F}$-weakly convergent at $P \in \mathcal{P}_0$, if there exists a family of probability measures $\{L(P,z), z \in \ell_\infty(\mathcal{F})\}$ on R such that

$$L_n(Q_n) \underset{n\to\infty}{\longrightarrow} L(P,z) \quad \text{weakly}$$

for every $z \in \ell_\infty(\mathcal{F})$ and every sequence $\{Q_n\} \subset \mathcal{P}_0$ such that

$$||n^{1/2}(Q_n - P) - z||_{\mathcal{F}} \underset{n\to\infty}{\longrightarrow} 0.$$

Remark 3.1. If the model is parametric, e.g., $P = P_\theta$, where $\theta = (\theta_1(P),\ldots,\theta_d(P))$ then the notion of locally $\mathcal{F}$-weakly convergent statistics reduces to the notion of a 'LAWC' statistic introduced by Beran (1992) when taking $\mathcal{F} = \{\theta_1,\ldots,\theta_d\}$.

Remark 3.2. If $\mathcal{F}$ is P-Donsker and if the statistic T_n is locally $\mathcal{F}$-weakly convergent at P, then there exists a random element Z with values in $\ell_\infty(\mathcal{F})$ such that the random probability measure $L_n(\widehat{P}_n)$, converges in distribution to a random probability measure $L(P,Z)$. Indeed, since $\mathcal{F}$ is P-Donsker, the empirical processes $\nu_{P,n}$ converge in law in $\ell_\infty(\mathcal{F})$ to a Gaussian process $G_P = \{G_P(f), f \in \mathcal{F}\}$ with zero mean and covariance $EG_P(f)G_P(g) = Pfg - PfPg$, $\; f,g \in \mathcal{F}$. By a result of Dudley (1985) there is a probability space $(\bar\Omega,\bar\Sigma,\bar P)$ and perfect functions $g_n : (\bar\Omega,\bar\Sigma,\bar P) \to (\Omega,\Sigma,P)$ such that $\bar P \circ g_n = P$ and

$$||n^{1/2}(\widehat{P}_n \circ g_n - P) - G_P \circ g_0||_{\mathcal{F}} \underset{n\to\infty}{\overset{\mathcal{D}}{\longrightarrow}} 0.$$

Hence, there is a set $\bar{\Omega}_0 \subset \bar{\Omega}$ such that $\bar{P}(\bar{\Omega}_0) = 1$ and

$$||n^{1/2}(\widehat{P}_n \circ g_n(\omega) - P) - G_P \circ g_0(\omega)||_{\mathcal{F}} \underset{n\to\infty}{\longrightarrow} 0$$

for each $\omega \in \bar{\Omega}_0$. By the definition of $\mathcal{F}$-weakly convergent statistics we have

$$L_n(\widehat{P}_n \circ g_n(\omega)) \underset{n\to\infty}{\longrightarrow} L(P, G_P \circ g_0) \text{ weakly.}$$

Example 3.3. The fact that a characteristic function is real if and only if the corresponding distribution function is symmetric at 0 suggests the statistic

$$T_n(Z_1, \ldots, Z_n; Q) = \int_{-\infty}^{+\infty} \Big(n^{1/2} \Im c_{Q,n}(t) + a(t)\Big)^2 g(t)dt,$$

for testing symmetry. Here

$$c_{Q,n}(t) = \int_{-\infty}^{+\infty} e^{itx} d\widehat{Q}_n(x)$$

denotes the empirical characteristic function corresponding to the distribution Q, g is an integrable weight function, and $a(t)$ satisfies $\int_{-\infty}^{\infty} a^2(t)q(t)dt < \infty$. This statistic T_n is locally $\mathcal{F}$-weakly convergent at any symmetric distribution P, when the class $\mathcal{F}$ is chosen as $\mathcal{F} = \{x \to \cos tx, \quad t \in R\}$.

A parametric version of the following proposition is given in Beran (1997).

PROPOSITION 3.2. *Assume that $\mathcal{F}$ is P-Donsker and that the statistic T_n is locally $\mathcal{F}$-weakly convergent at P. Let $\{m(n), n \geq 1\}$ denote any sequence of positive integers such that $m(n) \to \infty$ and $m(n)/n \to 0$ as $n \to \infty$. Then $L_{m(n)}(\widehat{P}_n)$ is d-consistent in probability, where d is any metric metrizing the weak convergence.*

Proof. We have

$$||m^{1/2}(\widehat{P}_n \circ g_n - P) - 0||_{\mathcal{F}} \underset{n\to\infty}{\longrightarrow} 0 \text{ a.s.}$$

By Definition 3.1, $L_m(\widehat{P}_n \circ g_n(\omega))$ converges weakly to $L(P, 0)$ for almost all $\omega \in \bar{\Omega}$. This completes the proof. ∎

In this subsection, we investigate the stochastic equivalence of the random distances $d(L_m(\widehat{P}_n), L_n(P))$ and $d(L_m(\widehat{P}_n), L_{m/2}(\widehat{P}_n))$, where d is either the Kolmogorov or the bounded Lipschitz distance. A unified way to consider both distances is to consider the more general class of uniform distances over a class of measurable bounded functions $\mathcal{H}$, say. Define for distributions F, Q, the uniform distance

$$d_{\mathcal{H}}(F, Q) = \sup_{h\in\mathcal{H}} |Fh - Qh|.$$

Indeed, if $\mathcal{H}$ is chosen as the class of indicator functions $\mathrm{I}\{(-\infty, a]\}$, $a \in R$, $d_{\mathcal{H}}(F, Q)$ will coincide with the Kolmogorov distance. Taking $\mathcal{H}$ as the class of measurable functions $h : R \to R$ such that $\sup_a |h(a)| + \sup_{a \neq b} |h(a) - h(b)|/|a - b| \leq 1$, then $d_{\mathcal{H}}(F, Q)$ we get the bounded Lipschitz metric. Note that distances $d_{\mathcal{H}}$, where $\mathcal{H}$ consists of higher order smooth functions have been used as well for investigating the accuracy of the bootstrap. Write

$$\Delta_{\mathcal{H}m} = d_{\mathcal{H}}(L_m(\widehat{P}_n), L_n(P)), \quad \widehat{\Delta}_{\mathcal{H}m} = d_{\mathcal{H}}(L_m(\widehat{P}_n), L_{m/2}(\widehat{P}_n)).$$

THEOREM 3.3. *Suppose that the sequence of statistics satisfies the assumptions* (A), (B), (C), *and* (D), *which are stated and discussed below. Assume furthermore that* $\mathcal{F}$ *is a* P*-Donsker class, and that* $||\nu_{P,n}||_{\mathcal{F}}$ *is uniformly integrable. For sequences* $m \in n(1/2, \gamma)$ *we may choose norming sequences* $\tau_{mn} = (n/m)^{1/2}$ *and* $m^{1/2}$ *corresponding to* $\gamma = \infty$ *and* $0 \leq \gamma < \infty$ *such that for random variables* ξ, ξ_1, ξ_2 *and a constant* $c_1 > 0$ *which are defined in the proof in Section 4*

$$\tau_{mn}\big(\Delta_{\mathcal{H}m}, \widehat{\Delta}_{\mathcal{H}m}\big) \ \underset{n\to\infty}{\overset{\mathcal{D}}{\longrightarrow}} \ (\xi_\gamma, \eta_\gamma),$$

where (writing $c := 1 - 2^{-1/2}$, $d = 2^{1/2} - 1$,*)* $(\xi_\infty, \eta_\infty) = (1, c)\,\xi$; $(\xi_0, \eta_0) = (1, d)\,c_1$ *and* $(\xi_\gamma, \eta_\gamma) = \big(\xi_1, \xi_2\big)$ *for* $0 < \gamma < \infty$.

Thus Theorem 3.3 yields the stochastic equivalence of the random distances $\widehat{\Delta}_{\mathcal{H}m}$ and $\Delta_{\mathcal{H}m}$, as $n \to \infty$.

In order to formulate the assumptions (A)–(D), fix a distance d on the set $\mathcal{P}_0$ and, for given constants $c_0 > 0$ and $c_1 > 0$, consider the neighborhood $\mathcal{V}(P) \subset \mathcal{P}_0$ of P defined by

$$\mathcal{V}(P) = \{Q \in \mathcal{P}_0 : d(Q, P) \leq c_1, \ \ ||n^{1/2}(Q - P)|| \leq c_0\}.$$

The first assumption concerns the local $\mathcal{F}$-weak convergence property of the sequence of statistics. Roughly speaking, we assume that parameterized expansions for $L_n(Q)h$ hold uniformly in the neighborhood $\mathcal{V}(P)$. A parameterization will be given by the quantity $n^{1/2}(Q-P)$ considered as an element in $\ell_\infty(\mathcal{F})$. In many cases, $\mathcal{F}$ will consist of a finite number of functions only.

Assumption (A). For each $Q \in \mathcal{V}(P)$, there exist a set $\{L(Q, z), z \in \ell_\infty(\mathcal{F})\}$ of probability distributions on R and a set $\{\ell(Q, z), z \in \ell_\infty(\mathcal{F})\}$ of real valued functions on $\mathcal{H}$ such that for every $h \in \mathcal{H}$

$$\begin{aligned} L_n(Q)h = L(Q, n^{1/2}(Q - P))h + n^{-1/2}\ell(Q, n^{1/2}(Q - P))h \\ + R_n(Q, n^{1/2}(Q - P), h), \end{aligned}$$

where

$$\sup_{Q \in \mathcal{V}(P)} \sup_{h \in \mathcal{H}} R_n(Q, n^{1/2}(Q - P), h) = o(n^{-1/2}).$$

Furthermore, we assume first order smoothness for $L(Q,z)$ and a Lipschitz condition for $\ell(Q,z)$ as a function of $z \in \ell_\infty(\mathcal{F})$.

Assumption (B). For each $h \in \mathcal{H}$ and $Q \in \mathcal{V}(P)$ we have

$$L(Q,z)h = L(Q,0)h + L_1(Q,h)z + R(Q,h,z),$$

where $L_1(Q,h)$ is a bounded linear functional on $\ell_\infty(\mathcal{F})$ and

$$\sup_{Q\in\mathcal{V}(P)} \sup_{h\in\mathcal{H}} |R(Q,h,z)| \le c_2(P)||z||_{\mathcal{F}}^2.$$

Moreover, $\sup_{h\in\mathcal{H}} |L_1(P,h)| \le \infty$.

Assumption (C). There exists a constant $\kappa(P)$ such that

$$\sup_{Q\in\mathcal{V}(P)} \sup_{h\in\mathcal{H}} |\ell(Q,0)h - \ell(Q,z)h| \le \kappa(P)||z||_{\mathcal{F}}$$

and $\sup_{h\in\mathcal{H}} |\ell(P,0)h| < \infty$.

Finally, we need the continuity of the limiting distribution $L(P) = L(P,0)$ as well as the continuity of the function $\ell(P) = \ell(P,0)$ at P.

Assumption (D). $d_{\mathcal{H}}(L(\widehat{P}_n,0), L(P)) = O_P(n^{-1/2})$ and

$$\sup_{h\in\mathcal{H}} |\ell(\widehat{P}_n,0)h - \ell(P)h| = O_P(n^{-1/2}).$$

Remark 3.4. If in assumption (B) the first oder approximation vanishes, i.e. if $L_1(Q,h) = 0$, we need again a second order expansion term for $L(Q,z)$ defined on $z \in \ell_\infty(\mathcal{F})$ to discriminate between the choice of m verses $m/2$ although now at a lower level. Hence, assumption (B) should be replaced by the following:

Assumption (B'). For each $h \in \mathcal{H}$ and $Q \in \mathcal{V}(P)$ we have

$$L(Q,z)h = L(Q,0)h + L_2(Q,h)z^2 + R(Q,h,z),$$

where $L_2(Q,h)$ is a bounded bilinear functional on $\ell_\infty(\mathcal{F})$ and

$$\sup_{Q\in\mathcal{V}(P)} \sup_{h\in\mathcal{H}} |R(Q,h,z)| \le c_2(P)||z||_{\mathcal{F}}^3.$$

Moreover, $\sup_{h\in\mathcal{H}} |L_2(P,h)| \le \infty$.

Example 3.5. Here we continue example 3.3. Consider the uniform distances

$$\Delta_m = \sup_{r\geq 0} |L_m(\widehat{P}_n; r) - L_n(P; r)|, \quad \widehat{\Delta}_m = \sup_{r\geq 0} |L_m(\widehat{P}_n; r) - L_{m/2}(\widehat{P}_n; r)| \tag{3.1}$$

corresponding to the distribution function of the statistic T_n defined in Example 3.3. There we assume $a \neq 0$. Write $T_n(Q) = \phi_a(S_n + z_n)$, where $S_n(t) = n^{-1/2}\sum_{k=1}^n \eta_k(t)$, $\eta_k(t) = [\cos Z_k t - E\cos Z_k t]$, $\phi_a(x) = \int_{-\infty}^{\infty}(x(t) + a(t))^2 q(t)dt$ and, finally, $z_n(t) = n^{1/2}E\cos Z_1 t$. We consider S_n as a sum of i.i.d. random elements in the Hilbert space $L_2(R, q(t)dt)$. We obtain, by the general result given in Theorem 3.4 below, that the random distances Δ_m and $\widehat{\Delta}_m$ are stochastically equivalent as $n \to \infty$. Namely, if $m \in n(1/2, \gamma)$, then $\widehat{\Delta}_m/\Delta_m \xrightarrow{\mathcal{D}} \xi_\gamma$, where $\xi_\infty = 2^{1/2} - 1$, $\xi_0 = 1 - 2^{-1/2}$, and ξ_γ is a random variable for $0 < \gamma < \infty$. Moreover, we get $E_P\widehat{\Delta}_m/E_P\Delta_m \longrightarrow c_\gamma(P)$, where $c_\gamma(P)$ is a constant depending on γ and P.

Example 3.6. Let H be a separable Hilbert space with the norm $||\cdot||$ and inner product $(\cdot,\cdot)$. Consider random elements $X_1, \ldots, X_n, ..$ in H that are independent and identically distributed with distribution P. Assume that $EX_1 = 0$ and P is taken from a class $\mathcal{P}_0$, with the following properties: i) Q is non symmetric around zero, ii) $\int_H ||x||^4 Q(dx) < \infty$, and iii) the covariance operator V_Q of Q has at least 13 (counted with multiplicities) eigenvalues exceeding a given $\beta > 0$.

The eigenvalues of a positive operator $V : H \to H$ will be denoted by $\lambda_1(V) \geq \lambda_2(V) \geq \cdots$. It is well known (see, e.g., Gohberg and Krein (1967)) that $|\lambda_j(V_1) - \lambda_j(V_2)| \leq ||V_1 - V_2||$ for linear completely continuous positive operators V_1, V_2 on H. Let $||\cdot||_2$ denote the Hilbert–Schmidt operator norm. One easily checks that

$$E||V_{\widehat{P}_n} - V_P||_2^2 \leq cn^{-1}\int_H ||x||^4 P(dx).$$

It follows that, with probability tending to one, at least $d \geq 13$ eigenvalues of the covariance operator $V_{\widehat{P}_n}$ will exceed a number $\beta_0 > 0$. Hence, without loss of generality we can assume that $\mathcal{P}_0$ contains the empirical distribution $\widehat{P}_n$ of $P \in \mathcal{P}_0$.

For $Q \in \mathcal{P}_0$ and $a \in H,\ a \neq 0$, define

$$T_{n,a}(Z_1, \ldots, Z_n; Q) = n||n^{-1}\sum_{k=1}^n Z_k + a||^2.$$

Let $L_{n,a}(Q, r)$ denote the distribution function of this statistic. Furthermore, let $\mathcal{F}$ denote the class of functions on H given by $x_1(z) = (x, z)$ together with $x_2(z) = (x, z)^2$ indexed by $x \in H$ with $||x|| = 1$. Note that the evaluation at a point $y \in H$ defines an embedding $H \subset \ell_\infty(\mathcal{F})$, via $y(x_1) := x_1(y) = (x, y)$ and for $y \in H$ and $y(x_2) := x_2(y) = (x, y)^2$.

It is easy to verify that the statistic $T_{n,a}$ is locally $\mathcal{F}$-weakly convergent at any $P \in \mathcal{P}_0$ which has zero mean. We aim to prove the stochastic equivalence as $n \to \infty$ of the uniform errors Δ_m and $\widehat{\Delta}_m$ defined by (3.1).

Denote by μ_Q the mean zero Gaussian measure on H with covariance operator V_Q, and let $D^j(x)\mu_Q$ be the jth directional derivative of μ in the direction of x. Set $V_r(z) = \{x \in H : ||x+z||^2 < r\}$. For $P, Q \in \mathcal{P}_0$, consider the distance

$$d(Q,P) = \Big| \int_H ||x||^4 d[Q(x) - P(x)] \Big| + ||V_Q - V_P||.$$

Define $\mathcal{V}(P) = \{Q \in \mathcal{P} : d(P,Q) \le c_1, \quad n^{1/2}||Q-P||_{\mathcal{F}} \le c_2\}$. An inspection of the proof of Theorem 2.1 in Bentkus (1984) gives the following uniform expansions (compare with assumption (A)). For any $Q \in \mathcal{V}(P)$ we have

$$L_n(Q,r) = \mu_Q(V_r(a+z_n)) + \frac{1}{6} n^{-1/2} ED^3(Z_1)\mu_Q(V_r(a+z_n)) + R_n(Q,a,r), \tag{3.2}$$

where $z_n = n^{1/2}EZ_1$ and, for any $\varepsilon > 0$ and $c_0 > 0$, there exists a constant $c > 0$ such that

$$\sup_{Q\in\mathcal{V}(P)} \sup_{||a||\le c_0} \sup_{r\ge 0} R_n(Q,z,r) = cn^{-1+\varepsilon}.$$

Furthermore, note that, for $z \in H$, we have

$$\mu_Q(V_r(a+z)) = \mu_Q(V_r(a)) + D(z)\mu_Q(V_r(a)) + R(Q,z;r), \tag{3.3}$$

where

$$\sup_{Q\in\mathcal{V}(P)} \sup_{||a||\le c_0} \sup_{r\ge 0} |R(Q,z;r)| \le c_2(P)||z||^2_{\mathcal{F}}.$$

Indeed, by standard arguments (inversion formula, Lebesgue lemma) it easily follows that

$$D^\ell(z)\mu_Q(V_r(a)) = (2\pi)^{-1} \int_{-\infty}^{+\infty} \frac{\partial^l}{\partial \tau^l} \Big[(it)^{-1} e^{-itr} q_t(a+\tau z)\Big]\Big|_{\tau=0} dt, \tag{3.4}$$

where $q_t(a)$ is the characteristic function corresponding to the measure $\mu_Q(V_r(a))$. By elementary calculations one proves that

$$|D^2(z)\mu_Q(V_r(a))| \le C(\lambda_1(Q)\cdots\lambda_5(Q))^{-1/2} \max\{1, ||a||^2\}||z||^2. \tag{3.5}$$

Hence, (3.3) follows by Taylor's formula and (3.5). Note, that (3.3) implies assumption (B). By (3.4) (which is assumption (C)) we get

$$\begin{aligned} &|ED^3(Z_1)\mu_Q(V_r(a+z)) - ED^3(Z_1)\mu_Q(a)| \le \\ &C(\lambda_1(Q)\cdots\lambda_9(Q))^{-1/2} \max\{1, ||a||^4\}||z||E||Z||^3. \end{aligned} \tag{3.6}$$

Finally, collecting the bounds (3.2), (3.3), and (3.6) we have proved the following result.

THEOREM 3.4. *Suppose that $Q \in \mathcal{V}(P)$ and $\int xQ(dx) \neq 0$. Then, for any constants $c_0 > 0$ and $\varepsilon > 0$, there exists a constant $c > 0$ such that*

$$\begin{aligned} L_n(Q,r) &= \mu_Q(V_r(a)) + D(z_n)\mu_Q(V_r(a)) \\ &+ \frac{1}{6n^{1/2}} ED^3(Z_1)\mu_Q(V_r(a)) + R_{m,n}(r), \end{aligned}$$

where

$$\sup_{Q\in\mathcal{V}(P)} \sup_{||a||\le c_0} \sup_{r\ge 0} |R_{m,n}(r)| \le c(n^{-1+\varepsilon} + ||z_n||^3 + n^{-1/2}||z_n||).$$

Arguing as in Götze and Zitikis (1997) one can prove that

$$\sup_{||a||\le c_0} \sup_{r\ge 0} |\mu_{\widehat{P}_n}(V_r(a)) - \mu_P(V_r(a))| = O(n^{-1/2})$$

and

$$\sup_{||a||\le c_0} \sup_{r\ge 0} |n^{-1}\sum_{k=1}^{n} D^3(X_k)\mu_{\widehat{P}_n}(V_r(a)) - E_P D^3(X_1)\mu_P(V_r(a))| = O(n^{-1/2}).$$

Combining this with Theorem 3.4, we get

COROLLARY 3.5. *If the sequence $m \in n(1/2,\gamma)$, then*

$$\frac{\widehat{\Delta}_m}{\Delta_m} \xrightarrow[n\to\infty]{\mathcal{D}} \xi_\gamma,$$

where $\xi_0 \equiv 1$, $\xi_\infty \equiv 1 - 2^{-1/2}$, and ξ_γ is a random variable for $0 < \gamma \le \infty$.

Remark 3.7. We excluded the case $a = 0$ since here $D(z)\mu_Q(V_r(0)) = 0$ and $D^3(z)\mu_Q(V_r(0)) = 0$. Therefore here one needs to use asymptotic expansions for $L_n(Q;r)$ up to order $o(n^{-1}.)$ This would require some additional use of Cramér's condition of smoothness for the measures $Q \in \mathcal{V}(P)$.

4. Proofs

Proof of Example 1.1. Set $t_- = (t - \overline{X}_n m^{1/2})/\widehat{\sigma}$, $t_+ = (t + \overline{X}_n m^{1/2})/\widehat{\sigma}$, where, as usual, $\widehat{\sigma}^2$ denotes the sample variance. We also denote, for $s > 0$, $\kappa_s = EX_1^s$. By Edgeworth expansions we have

$$\begin{aligned} P(m(\overline{X}_m^*)^2 < t^2 \big| X_1,\dots,X_n) &= \Phi(t_-) + \Phi(t_+) - 1 + m^{-1/2}(\widehat{Q}_1(t_-) - \widehat{Q}_1(-t_+)) \\ &+ m^{-1}(\widehat{Q}_2(t_-) - \widehat{Q}_2(-t_+)) + o_P(m^{-1}), \end{aligned}$$

where

$$Q_1(t) = -\frac{e^{-t^2/2}}{\sqrt{2\pi}} H_2(t)\frac{\kappa_3}{6\sigma^3}, \quad Q_2(t) = -\frac{e^{-t^2/2}}{\sqrt{2\pi}}\Big[\frac{\kappa_4}{24\sigma^4}H_3(t) + \frac{\kappa_3^2}{72\sigma^6}H_5(t)\Big]$$

and $\widehat{Q}_1$ and $\widehat{Q}_2$ denote the plug-in estimates of Q_1 and Q_2 respectively. Here H_s denote the standard Hermite polynomials. Straightforward calculations show that

$$
\begin{aligned}
&P(m(\overline{X}_m^*)^2 < t^2 \big| X_1, \dots, X_n) \\
&= 2\Phi(t/\widehat{\sigma}) - 1 + \frac{m}{n}\Phi''(t/\widehat{\sigma})Y_n^2 \\
&\quad - \frac{2}{n^{1/2}}\widehat{Q}_1'(t/\widehat{\sigma})Y_n + \frac{2}{m}\widehat{Q}_2(t/\widehat{\sigma}) + \mathcal{O}_P\Big(\Big(\frac{m}{n}\Big)^{3/2} + \frac{1}{n}\Big) + o_P\Big(\frac{1}{m}\Big),
\end{aligned} \tag{4.1}
$$

where $Y_n = (n^{1/2}\overline{X}_n)/\widehat{\sigma}$. Furthermore,

$$
P\left(n(\overline{X}_n)^2 < t^2\right) = 2\Phi(t/\sigma) - 1 + o(n^{-1/2}). \tag{4.2}
$$

Set

$$
\zeta_m(t) := P\left(m(\overline{X}_m^*)^2 < t^2 \big| X_1, \dots, X_n\right) - P\left(n(\overline{X}_n)^2 < t^2\right)
$$

and $\widehat{\zeta}_m(t) = \zeta_m(t) - \zeta_{m/2}(t)$. If $n/m^2 \to 0$, then (4.1) and (4.2) together yield

$$
\frac{n}{m}\sup_{t\ge 0}|\zeta_m(t)| = Y_n^2 \sup_{t\ge 0}|\Phi''(t)| + o_P(1)
$$

and

$$
\frac{n}{m}\sup_{t\ge 0}|\widehat{\zeta}_m(t)| = 2^{-1}Y_n^2 \sup_{t\ge 0}|\Phi''(t)| + o_P(1).
$$

Combining this with the central limit theorem we get (1.2). The proof of (1.3) immediately follows by (4.1), (4.2), and the law of large numbers. For the proof of (1.4), we have, by (4.1), (4.2) combined with the law of large numbers,

$$
\begin{aligned}
n^{1/2}\Delta_m &= \sup_{t\ge 0}|\zeta_m(t)| \\
&= \sup_{t\ge 0}|(m/n^{1/2})\Phi''(t)Y_n^2 - 2Q_1'(t)Y_n + 2(n^{1/2}/m)Q_2(t)| + o_P(1)
\end{aligned}
$$

and

$$
\begin{aligned}
n^{1/2}\widehat{\Delta}_m &= \sup_{t\ge 0}|\widehat{\zeta}_m(t)| = \\
&= \sup_{t\ge 0}|(m/2n^{1/2})\Phi''(t)Y_n^2 - 2(n^{1/2}/m)Q_2(t)| + o_P(1).
\end{aligned}
$$

It is not difficult to verify that (1.5) is valid with $f_1(Y) = \sup_{t\ge 0}|\gamma\Phi''(t)Y^2 - 2Q_1'(t)Y + 2\gamma^{-1}Q_2(t)|$ and $f_2(Y) = \sup_{t\ge 0}|2^{-1}\gamma\Phi''(t)Y^2 - 2\gamma^{-1}Q_2(t)|$. ∎

The proof of Theorem 2.1 is using two results about U-statistics due to Vitale (1992) which we state below. For a sequence of i.i.d. random elements $X_1, \dots, X_n$ taking values in a measurable space $(S, \mathcal{A})$ and a sequence of functions (h_m), where $h_m : S^m \to R$ is a real-valued kernel of degree $m \le n$, define the U-statistic $U_{n,m}h_m$ by

$$
U_{n,m}h_m = \binom{n}{m}^{-1} \sum_{1\le i_1 < \cdots < i_m \le n} h_m(X_{i_1}, \dots, X_{i_m}).
$$

whereas the conditional kernel $h_{m|k} : S^k \to R$ is defined by

$$h_{m|k}(x_1, \dots, x_k) = \int_S \cdots \int_S h_m(x_1, \dots, x_m) P(dx_{k+1}) \cdots P(dx_m).$$

Then the Hoeffding decomposition of $U_{n,m}h_m$ is given by

$$U_{n,m}h_m - EU_{n,m}h_m = \sum_{k=1}^{m} \binom{m}{k} U_{n,k}h_{m(k)},$$

where

$$h_{m(k)}(x_1, \dots, x_k) = \sum_{j=0}^{k} (-1)^{k-j} \sum_{1 \le i_1 < \dots < i_j \le n} h_{m|j}(x_{i_1}, \dots, x_{i_j}).$$

The degree of degeneracy of the U-statistics $U_{n,m}h_m$ is the largest integer r such that provided $h_{m(k)} = 0$ for $k = 1, \dots, r$.

LEMMA 4.1. *Suppose that $U_{n,m}h_m$ is the U-statistics based on the symmetric kernel h_m satisfying $Eh_m^2 < \infty$. If the degree of degeneracy of U_n is equal to $r-1$, then*

$$\operatorname{var}(U_n) \le \frac{\binom{m}{r}^2}{\binom{n}{r}} \operatorname{var}(h_{m|r}) + \frac{\binom{m}{r+1}}{\binom{n}{r+1}} \Big(\operatorname{var}(h_m) - \binom{m}{r} \operatorname{var}(h_{m|r}) \Big).$$

LEMMA 4.2. *If $U_{n,m}h_m$ is as in Lemma 4.1, the sequence $\operatorname{var}(h_{m|k})/\binom{k}{r}$, $k = r, \dots, m$ is nondecreasing.*

Proof of Theorem 2.1. Write

$$\zeta^*_{n,m}(a) = U_{nm}(a) + V_{nm}(a) + R_{nm}(a),$$

where $U_{nm}(a)$ is U-statistic with kernel

$$h_m(x_1, \dots, x_m; a) = h(T_m(x_1, \dots, x_m; P); a) - \theta_m(P; a);$$

$V_{nm}(a) = \theta_m(P;a) - \theta_n(P;a)$, $R_{nm}(a) = \theta_{mn}(\widehat{P}_n;a) - \theta_{mn}(P;a)$, $a \in R$. By assumption (J) $||R_{nm}||_2 = o_P((m/n)^{1/2} + m^{-1/2})$ whereas assumption (I) yields

$$V_{nm}(a) = \frac{1}{\sqrt{m}} E\kappa(X_1; a) + v_{nm}(a), \quad a \in R,$$

where $||v_{nm}||_2 = o(m^{-1/2})$. Hoeffding's decomposition yields

$$U_{nm}(a) = \frac{m}{n} \sum_{i=1}^{n} g_m(X_i; a) + r_{n,m}(a),$$

where

$$g_m(x; a) = E\big(h(T_m; a) \big| X_1 = x\big) - \theta_m(P; a)$$

and, in view of Lemma 4.1 and Lemma 4.2, we obtain

$$Er_{nm}^2(a) \le \frac{m^2}{n^2} E\big(h(T_m; a) - \theta_m(P; a)\big)^2.$$

This gives $||r_{nm}||_2 = o_P((m/n)^{1/2} + m^{-1/2})$. Finally, assumption (I) leads to

$$\zeta_{n,m}^*(a) = \frac{\sqrt{m}}{n} \sum_{k=1}^n (\kappa(X_k; a) - E\kappa(X_k; a)) + \frac{1}{\sqrt{m}} E\kappa(X_1; a) + \kappa_{n,m}(a), \tag{4.3}$$

for every $a \in R$, where $||\kappa_{m,n}||_2 = o_P((m/n)^{1/2} + 1/m^{1/2})$. Now the result easily follows from the representation (4.3) and the central limit theorem in Hilbert spaces. ∎

Proof of Theorem 2.3. Under the assumption (J') the L_2-weak limiting behavior of the process $\tau_{mn}\zeta_m^*$ coincides with that of the process $\tau_{mn}U_{nm}$, where U_{mn} denotes the L_2-valued U-statistic, $U_{mn}(a) = \theta_{mn}(P; a) - \theta_n(P; a),\ a \in R$. Hoefding's decomposition, Lemma 4.1, Lemma 4.2 and assumption (I_2) reduce the proof to the weak convergence of the L_2-valued random elements

$$S_{mn}(a) = \frac{m}{n} \sum_{k=1}^n g_m(X_k; a) + \frac{1}{m}\beta_P(a),\ a \in R.$$

This can be easily shown using the results of Cremers and Kadelka (1986) on weak convergence in L_p spaces. ∎

Proof of Theorem 3.3. Introduce $\tau_{m,n} = (m/n)^{1/2} + m^{-1/2}$ and fixing constants $c_0 > 0, c_1 > 0$, define the set $\Omega_0 = \{||\nu_{P,n}||_{\mathcal{F}} \le c_0(n/m)^{1/2}\} \cap \{d(P_n, P) \le c_1\}$. On this set Ω_0, we have by assumption (A)

$$\begin{aligned} H(h) := L_m(\widehat{P}_n)h - L_n(P)h &= L(\widehat{P}_n, m^{1/2}(\widehat{P}_n - P))h - L(P,0)h + \\ + &\quad \frac{1}{m^{1/2}}\ell(\widehat{P}_n, m^{1/2}(\widehat{P}_n - P))h - \frac{1}{n^{1/2}}\ell(P,0)h + R_{m,n}(h), \end{aligned}$$

where $\sup_{h\in\mathcal{H}} |R_{m,n}(h)| = o_P(m^{-1/2})$.

Write $\widehat{a}_n(h) = L_1(\widehat{P}_n, h)(\nu_{P,n})$, $\widehat{b}_n(h) = \ell(\widehat{P}_n, 0)h$. Using with assumptions (B) and (C) we get

$$H(h) = (m/n)^{1/2}\widehat{a}_n(h) + m^{-1/2}\widehat{b}_n(h) + T_{mn}(h),$$

where $\sup_{h\in\mathcal{H}} |T_{mn}(h)| = o_P(\tau_{mn})$. In view of assumption (D), we conclude

$$H(h) = (m/n)^{1/2} a_n(h) + m^{-1/2} b(h) + T_{mn}(h), \tag{4.4}$$

where

$$a_n(h) = L_1(P, h)(\nu_{P,n}), \quad b(h) = \ell(P,0)h.$$

Now (4.4) yields

$$H^*(h) := L_m(\widehat{P}_n)h - L_{m/2}(\widehat{P}_n)h = \qquad (4.5)$$
$$= (1-2^{-1/2})(m/n)^{1/2}a_n(h) - (\sqrt{2}-1)m^{-1/2}b(h) + T^*_{m,n}(h),$$

where $\sup_{h\in\mathcal{H}}|T^*_{m,n}(h)| = o_P(\tau_{mn})$. Since $\mathcal{F}$ is P-Donsker and $||\nu_{P,n}||_{\mathcal{F}}$ is uniformly integrable, we obtain

$$\sup_{h\in\mathcal{H}}|a_n(h)| \xrightarrow[n\to\infty]{\mathcal{D}} \sup_{h\in\mathcal{H}}|L_1(P,h)G_P|. \qquad (4.6)$$

Using (4.4), (4.5), and (4.6) it is easy to complete the proof of the theorem. Moreover, we observe that $\xi = \sup_{h\in\mathcal{H}}|L_1(P,h)G_P|$, $c_1 = \sup_{h\in\mathcal{H}}|b(h)|$, whereas

$$\xi_1 = \sup_{h\in\mathcal{H}}|\gamma L_1(P,h)G_P + b(h)| \quad \text{and}$$
$$\xi_2 = \sup_{h\in\mathcal{H}}|\gamma(1-2^{-1/2})L_1(P,h)G_P + (1-\sqrt{2})b(h)|.$$

■

REFERENCES

Arcones, M.A. and Gine, E., (1989). The bootstrap of the mean with arbitrary bootstrap sample size. *Ann. Inst. Henri Poincare* **25**, 457–481.

Athreya, A.C., (1987). Bootstrap of the mean in the infinite variance case. *Annals of Statistics* **15**, 724–731.

Bentkus, V., (1984). Asymptotic expansions for distributions of sums of independent random elements of a Hilbert space. *Lithuan. Math. J.* **24**, 305–319.

Bentkus, V. and Götze, F., (1999). Optimal bounds in non-Gaussian limit theorems for U-statistics. *Ann. Probab.* bf 27, 454–521.

Beran, R., (1982). Estimated sampling distributions: the bootstrap and competitors. *Ann. Statist.* **10**, 212–225.

Bickel, P.J., Götze, F. and Zwet, W.R., (1997). Resampling fewer than n observations: gains, losses, and remedies for losses. *Statistica Sinica* **7**, 1–32.

Bickel, P. J. and Freedman D. A., (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.* **9**, 1196–1217.

Bretagnolle, J., (1983). Lois limites du bootstrap de certaines fonctionelles. *Annales Inst. Henri Poincare* **19**, 281–296.

Cörgő, S. and Mason, D. (1990). Bootstrapping empirical functions. *Ann. Statist.* **17**, 1447–1471.

Cremers, H. and Kadelka, D. (1986). On weak convergence of integral functionals of stochastic processes with applications to processes taking paths in L_p^E. *Stoch. Proc. Appl.* **21**, 305–317.

Davison, A.C. and Hinkley, D.V., (1998). *Bootstrap Methods and Their Applications* Cambridge University Press, United Kingdom, New York, Melbourne.

Datta, S., (1996). On asymptotic properties of bootstrap for $AR(1)$ processes. *J. Stat. Plan. Infer.* **53**, 361–374.

Datta, S. and McCormik, W.P., (1995). Bootstrap inference for a first order autoregression with positive innovations. *JASA* **90**, 1289–1300.

Deheuvels, P., Mason, D. and Shorack, G., (1993). Some results on the influence of extremes on the bootstrap. *Ann. Inst. H. Poincare* **29**, 83–103.

Dudley, R.M., (1984). A course on empirical processes. *Lect. Notes in Math. Springer, Berlin* **1097**, 1–142.

Dudley, R.M., (1985). An extended Wichura theorem, definition of Donsker class and weighted empirical distributions. *Lect. Notes in Math. Springer, Berlin* **1153**, 141–178.

Gohberg, I. C. and Krein, M. G., (1969). *Introduction to the Theory of Linear Nonselfadjoint Operators* American Mathematical Society, Providence.

Götze, F., (1993). *IMS Bulletin*

Götze, F., (1985). Asymptotic expansions in functional limit theorems. *J. Multivariate Anal.* **16**, 1–20 .

Götze, F. and Zitikis, R. (1995). Edgeworth expansions and bootstrap for degenerate von Misesw statistics. *Prob. Math. Stat.* **15**, 327–351.

Hall, P., (1992). *The Bootstrap and Edgeworth Expansion* Springer, New York.

Heimann, G. and Kreiss, J.P., (1996). Bootstrapping general first order autoregression. *Stat. Probab. Letters* **30**, 87–98.

Politis, D. N. and Romano, J. P., (1994). A general theory for large sample confidence regions based on subsamples under minimal assumptions. *Ann. Statist.* **22**, 2031–2050.

Shao, J., (1994). Bootstrap sample size in nonregular cases. *Proc. Amer. Math. Soc.* **122**, 1251–1262.

Shao, J., (1996). Bootstrap model selection. *JASA* **91**, 655-665.

Swanepoel, J. W. H., (1986). A note on proving that the (modified) bootstrap works. *Comm. Statist. Theory Methods* **15**, 3193–3203.

Department of Mathematics
Bielefeld University
Postfach 100131
33501 Bielefeld
Germany
goetze@mathematik.uni-bielefeld.de

Department of Mathematics
Vilnius University
Naugarduko 24
2600 Vilnius, Lithuania
alfredas.rackauskas@maf.vu.lt